

UNITED STATES PATENT & TRADEMARK OFFICE

Application for

**“System and Method for Processing Formatted
Text Documents in a Database”**

Inventors:

Forrest Rhoads

Dan Dabney

Bob Haschart

Jack Conrad

Howard Turtle

Kinley Larntz

Bokyoung Yang

Dan Gannon

"Express Mail" mailing label number: EL600376552US

Date of Deposit: December 22, 2000

This paper or fee is being deposited on the date indicated above with the United States Postal Service pursuant to 37 CFR 1.10, and is addressed to the Commissioner for Patents, Box Patent Application, Washington, D.C. 20231.

SYSTEM AND METHOD FOR PROCESSING FORMATTED
TEXT DOCUMENTS IN A DATABASE

5 Background of the Invention

This invention relates generally to a system and method for processing a plurality of textual documents and in particular to a system and method for identifying characteristics within the textual documents and for extracting relationships between the textual documents
10 based on the identified characteristics.

The explosion in the number of textual documents being generated has made it increasingly important to generate an electronic version of the documents to enable automated processing to extract data, to determine information about the textual documents,
15 and to identify relationships with other textual documents in a database. This is especially true with very large databases which may contain hundreds of thousands of textual documents, such as a database containing legal cases and other legal material.

A legal database may contain a large number of legal cases. A legal case in this
20 document refers to an individual written decision issued in the course of a litigation. These decisions usually contain citations to and quotations from other documents, including other legal cases to establish past practice and justify the result (establish precedence) of the decision. Citations are written in distinctive styles which includes special abbreviations and

punctuation which facilitate their identification. Also, quotations are usually set off in quotation mark. It is useful to identify these citations and quotations.

In some circumstances, the citations and quotations in a legal case may be identified by automatically parsing through the text of a legal case to identify candidates based upon punctuation and other characteristics. However, the punctuation which sets off a quotation may be used for other purposes, and the abbreviations and formats which characterize a citation are not necessarily unique. To ensure accurate identification, a citation or quotation must be verified. A citation has a predefined format. For example, a written decision of a California Court of Appeals case in 1993 in a lawsuit between Ms. Pleasant and Mr. Celli may have a citation such as, *Pleasant v. Celli*, 18 Cal. App. 4th 841, 22 Cal. Rptr. 2d 663 (1993). The first portion of the citation indicates the two parties last names and their positions in the case. For example, Ms. Pleasant is the first name listed. Since this is an appellate case, she is the appellant and Mr. Celli is the appellee. The next portion of the citation (i.e., 18 Cal. App. 4th 841 for example) indicates where a copy of the written decision may be located. The first number indicates the volume number of the case reporter in which the decision is located and the text portion (i.e., Cal. App 4th) indicates the name of the reporter and the edition of the reporter. The number following the reporter name indicates the starting page number of the decision. Thus, both the text of the citations and quotations in a written reported decision textual document have well defined format which may be automatically identified.

Once the citations and quotations in a legal case are identified and any relationships between these citations and quotations and any prior legal cases are determined, this information may be used for a variety of purposes. For example, this information may be used for both legal case verification purposes and legal case collocation purposes.

- 5 Verification is a process whose end result is a determination that the legal case currently being reviewed is still good law (i.e., it has not been overruled or limited by some later case due to different reasoning). A case that is not good law may not be persuasive since the reasoning of the case is no longer valid. Thus, the process of verification ensures that the case being used to support an argument is still good law and the reasoning of that case is still
- 10 valid. Collocation, on the other hand, is a process whose end result may be a list of other legal cases, legal materials or textual documents which cover similar issues to the case currently being reviewed, or a list of cases covering a particular subject matter, such as intellectual property. For example, a user of the legal database may have located a case which is of interest to him and he would then like to identify other cases that are related to
- 15 the case of interest based on the subject matter. Thus, it is desirable for a user of a legal database to be able to perform both verification and collocation using a single integrated system.

Prior to electronic databases, legal cases were published in several different formats,

20 each of which have some advantages and disadvantages. A register publication gathers information at the source of the new law and presents the new law in roughly chronological order. An example of this type of publication is a reporter volume that contains the new legal cases for a particular court in chronological order. These register publications may be

rapidly published and provide a good statement of the new law, but these register publications cannot provide collocation (i.e., it cannot provide a researcher with information about other cases which may be related to the subject matter of the instant case). It is also difficult to verify a legal case with a register publication. Another type of legal publication is a code publication which attempts to gather together the law that applies to a particular subject matter so that a researcher may determine the current state of the law from a single source. A code publication may permit a case to be verified since outdated law is removed as a code publication is updated. An example of a code publication is an annotated statutory code publication which gathers information from statutes and legal cases about a particular subject matter. Legal case are not published in a code format. These code publications may provide a researcher with the ability to collocate legal cases and verify a legal case, but the information in these code publications, due to the amount of time required to compile these code publications, only describe the state of the law sometime in the recent past. Therefore, for rapidly changing areas of the law, a code publication may not be accurate and therefore cannot provide either the verification or collocation needs to the researcher.

Numerous efforts have been made to codify case law, but case law tends to resist codification for several reasons. First, there is a tremendously large number of legal cases so it is often very difficult and time consuming to attempt to classify the legal cases by subject matter. In addition, legal cases have a complex hierarchy of authority and control (i.e., whether a case from a particular court controls or influences the outcome of another court, which leads to differences between various courts within various regions of the United States. It is also very difficult to determine the effects that a newly-decided legal case may

have on the current state of the law. The attempts to codify the case law have included treatises (which cover the current state of the law for a particular subject matter) or restatements of the law in particular subject areas.

5 In order to adequately perform legal research, legal researchers need to be able to determine the current state of the case law, and then access the case law in its register form. Any access to the case law itself or free-text access cannot easily provide a current state of the law. Thus, a researcher may use an index to the case law to find the case law. These indexes include a controlled vocabulary, manually created intellectual index to case law
10 which requires extraordinary amounts of manpower to implement. Instead of these options, a citation index (also known as a case law citator) may be used.

There are several conventional case law citators. One citator provides both the required verification and collocation functions, but has several disadvantages. In particular,
15 this citator produces citation chains for a particular legal case which list how the legal case was treated by later legal case, such as indicating that the particular legal case was overruled or followed. This citator is not current since it typically waited for a legal case to be in the general case law prior to generating the case law citator information. This conventional system did an adequate job of verification, but did not really provide an adequate collocation
20 function. In particular, it is tedious to locate other case law which relates to a particular case by following the citation strings, retrieving the reporter with a legal case, and then manually reviewing the legal case.

Therefore, it is desirable to provide a system and method which seamlessly integrates general subject matter access to case law with citation information which provides a researcher with the desired verification function as well as the collocation function. Thus, the researcher may use a single integrated system which provides the researcher will all of the information that the researcher needs in a single location. The invention provides such a system, as described below.

Summary of the Invention

The invention provides a system which processes a textual document, such as a legal case, automatically generates information about the textual document, and automatically establishes relationships between the textual document and other textual documents in a database based on the generated information. The system then permits a user of the system to select a textual document and view any other textual documents which have a particular relationship to the selected textual document based on citations in the selected text document.

This invention provides some automated processes which make it possible to create an index both more quickly and more economically. In addition, the system provides a user with a richer characterization of each citation because it provides information about the citation in a graphical manner.

In one aspect, the invention provides a legal database system that uses an automatic text processing system to permit a researcher to verify a legal case as well as to collocate other legal cases. The automatic text processing system processes each legal case to generate information about the legal case, and then establishes relationships between the legal case and other legal cases based on certain criteria. To provide these functions in a efficient, cost effective manner, the automatic text processing system may identify and verify quotations in the written reported decision of the legal case, generate information about “negative treatment” of the legal case by other legal cases (e.g., the reasoning of a written opinion of a legal case may be disagreed with, overruled or distinguished by the written decisions of other legal cases), generate “depth of treatment” information (e.g., the significance of a citation in a written decision of the legal case based on, for example, how much effort and words were spent discussing the cited case), and generate a subject matter classification for each citation in the legal case. The system reduces the time necessary to process a legal case and reduces the cost to provide case law citation information yet maintains the same accuracy as a conventional system. The invention also parses a textual document, identifies any quotations in the textual document and verifies the source of the quotation. The invention also determines a subject matter classifications for each citation identified in a legal case.

In accordance with the invention, a system for automatically determining a plurality of pieces of different information about a textual document from the text of the textual document is provided. The system may generate a data record containing the identified characteristics of the textual document, the negative treatment of other textual documents

cited in the textual document, the depth of treatment of there textual documents cited in the textual document, the verified quotation information, and the belief scores of the top-ranked headnotes so that all of the information is accessible. To generate the data record, one or more particular characteristics of the text of the textual document are identified to generate a file containing the text of the textual document with the particular characteristics marked up within the text. Any likely negative treatment of other textual documents from the text of the textual document may be identified to generate negative treatment information, a data record is produced indicating a depth of treatment for each other textual documents cited in the text of the textual document, the source of a quotation in the text of the textual document is verified to generate verified quotation information, and any subject matter text that is applicable to the text of the textual document is determined to generate subject matter headings (e.g., headnotes) associated with the citations in the text of the textual document.

In another aspect, the invention provides a method for processing a textual document to identify quotations within the textual document in which the text of the textual document is parsed into one or more paragraphs and one or more symbols are identified in each paragraph. The symbols, such as quotation marks, indicate the presence of a quotation in the sentence, and quotations within the text of the textual document are determined based on the identified tokens. In addition, a source of the quotation is verified by matching the text of the quotation with the text of the actual quotation within the textual document, and a verified quotation data record is generated.

Brief Description of the Drawings

Figure 1 is a block diagram of a computer system which may be used to automatically generate case law citation data in a legal text processing system in accordance with the invention;

Figures 2A - 2D are screen shots illustrating examples of the user interface displayed to a user of the case law citation system in accordance with the invention;

Figure 3 is a diagram illustrating the flow of data in a case law citation system in accordance with the invention;

Figure 4 is a diagram illustrating more details of the flow of data in the case law citation system of Figure 3.

Figure 5 is a diagram illustrating a system for quotation identification in accordance with the invention that may be part of the case law citation system;

Figure 6 is a flowchart illustrating a method for quotation identification and verification process in accordance with the invention;

Figure 7 is a flowchart illustrating a method for quotation verification in accordance with the invention;

Figure 8 is a diagram illustrating a negative history determining process in accordance with the invention;

Figure 9 is a flowchart illustrating a method for determining the depth of treatment of a legal case in accordance with the invention; and

Figure 10 is a diagram illustrating a method for assigning subject matter classifications in accordance with the invention.

Detailed Description of a Preferred Embodiment

The invention is particularly applicable to a computer-implemented legal text processing system and method for semi-automatically identifying characteristics, such as citations and quotations, within a legal document and identifying relationships between the legal document and other legal documents stored in the database. The legal document may be a legal case, a statute, a law review article, an ALR article or a legal treatise. It is in the context of a legal case that the invention will be described. It will be appreciated, however, that the system and method in accordance with the invention has greater utility and may be used for different legal documents.

Before describing the preferred embodiment of the invention, a brief description of the terminology that will be used to describe the invention will be provided. Any reported decision of a legal case is presumed to be an authoritative statement of the law when it is written. Then, later events may affect the authoritativeness of this legal case's decision. These later events may include later proceedings or written decisions during the same litigation (e.g., direct history), a decision of a later legal case from a different litigation which resolves the same issues in a different way or using different reasoning and overrules the earlier case, or a decision of a later legal case from a different litigation which resolves the same issue differently, but does not explicitly overrule the case. The direct history of a legal case may include a record of the connections between the legal cases that are part of the same litigation. The direct history may be of varying degrees of relevance and may include

positive history (i.e., maintaining or supporting the authority of the legal case) or negative history (for example, the legal case may no longer have the authority it once had). The indirect history of a case is a record of the connections between legal case and other legal cases which are not part of the same litigation. The indirect history of a legal case may also be positive or negative. The significance of a particular case may often be indicated by the amount of discussion (i.e., the amount of text) that a later case uses in discussing a decision of another legal case while following, overruling or explaining the case. This is referred to as the depth of treatment of the case, as described below. Now, a preferred embodiment of the invention will be described.

Figure 1 is a block diagram of a computer system 30 in which the invention may be embodied. The system may semi-automatically identify characteristics, such as citations and quotations within a legal case document, and then generate information about the legal case in the context of other legal cases. The computer system may include a computer 32, a server 34 and a plurality of client computers 36. The computer 32 may further include a central processing unit (CPU) 38, a memory 40 and one or more processes 42, which may be software applications that are stored in the memory 40. The CPU controls the operation of the computer and executes the software applications stored in the memory. In operation, a plurality of pieces of electronic data corresponding to the text of the published decisions for the legal cases are fed into the computer and temporarily stored in the memory 40. In the following discussion, the written opinion of the legal case is referred to as the legal case. Each piece of electronic data (i.e., each written opinion of a legal case) may be automatically processed by the CPU, using the processes contained in the software applications contained

in the memory, to generate information about the legal case, as described below. For example, the CPU may parse the text of the legal case to identify candidate (i.e., unverified) citations to other legal cases and mark these citations for later processing, may identify candidate (i.e., unverified) quotations in the text of the legal case and mark the text accordingly, may verify the source of a quotation in the text of the legal case, may determine a depth of treatment of a cited legal case (i.e., the significance of the cited legal case based on some predetermined criteria), may determine the negative treatment of the legal case, and may assign subject matter text, such as headnotes, in accordance with a predetermined classification system to citations in the legal case. Each of these processes may be performed by a software application in the memory 40, which is executed by the CPU 38. The details of each of these processes will be described below.

Once the processing has been completed by the processes 42, the computer 32 outputs a data record 44 for the particular legal case which contains information about the history of the legal case, information about the depth of treatment of citations in the legal case, information about quotations within the legal case, and information about the subject matter text (i.e., headnotes) assigned to each citation in the legal case. The data record generated by the computer 32 for each legal case may be stored in a database 33 in a server 34. Then, when a user of one of the plurality of client computers 36 requests information about a legal case, the server 34 generates a user interface containing a variety of information about the requested legal case based on the data records in the database 33, and presents the user reviewing the legal case with a variety of information about the legal case. An example of the user interface provided to the user of each client computer is described

below with reference to Figures 2A - 2D. In this manner, a user of the client computer may request data about a particular case, and the system in accordance with the invention provides that data to the user.

5 As the electronic data for the text of each written opinion for a new legal case is received by the computer 32, the legal case is processed as described above and the results of the processing is stored as a data record 44 in the database 33 of the in the server 34. The users of the client computers may then retrieve data about a particular legal case from the server 34. Thus, while the server 34 is providing data about a legal case to the one or more
10 users of the client computers, the computer 32 may be simultaneously processing additional new legal cases and adding the information for that new legal case into the database 33 in the server 34. Now, an example of a preferred user interface and information provided to the user of a client computer will be described in more detail.

15 Figures 2A - 2D are screen shots illustrating examples of a preferred user interface and the information provided to the user of a client computer in accordance with the invention. Figure 2A shows a computer screen 50 on a client computer displaying a legal case being reviewed by the user of the particular client computer in which the user interface has a Windows format, a toolbar, pull down menus, etc. In this example, the display is of
20 the text of a legal case called *Pleasant v. Celli* which was decided by a California Court of Appeals. As described above, any citation for a legal case has a well-defined format which facilitate the identification of these citations within the text of the written opinion of the legal case. In order to access more information about the displayed legal case, the user of

the client computer may select the citation service, which may be referred to as KeyCite™, from the Services menu 51 by clicking on a “KC” button 52 or click on a symbol 54.

KeyCite™ is a trademark of a citator of the assignee of the present invention. The symbol may be a colored symbol, e.g., a flag, which gives a quick status of the legal case. A red colored flag may warn that the legal case being reviewed may not be good law for at least some portion of the legal case, a yellow colored flag may indicate that the legal case has some negative history, as described below, or another colored symbol, such as a blue H, may indicate that the legal case has some history which is not negative. The invention, however is not limited to any particular types of symbols or colors. Once the user of the client computer has selected the citator system in some manner, the screen shown in Figure 2B may be displayed.

Figure 2B is a screen shot showing an example of a computer screen 50 which the invention may employ having a control interface portion 58, and a display portion 60. The control interface portion of the display permits the user to customize the information being displayed. For example, if a first radio button 62 is selected, then the full history of the legal case, including direct history which is negative or positive, negative indirect history, and any related references may be shown. If a second radio button 64 is selected, then only the negative direct and indirect history may be shown. If a third radio button 66 is selected, then only the direct history of the legal case may be displayed so that any minor direct history (including references), remote direct history (such as appeals after remand) and mildly negative indirect history are not displayed. The control portion 58 of the display also may indicate the number of cases which are considered to be the history of the legal case. The

control portion 58 may also include a fourth radio button 68 and an indication of the number of citations to the legal case being displayed. When the fourth radio button is selected, a list of other documents is displayed.

5 In the example shown in Figure 2B, the full history of the *Pleasant v. Celli* case is indicated. As shown, the various types of history, such as the direct history and the negative indirect history are displayed in the display portion 60 and are separated from one another by headings. For each piece of history, a short description of the history or tag, such as “opinion vacated by”, “disapproved of by”, or “disagreed with by” may indicate the
10 relationship between the cases listed and the base case. In this example, an earlier decision of the same court was vacated by the *Pleasant* case. Now, the citations to the legal case will be described with reference to Figure 2C.

Figure 2C is an example of a screen shot showing the computer screen 50 having the
15 control interface portion 58, and the display portion 60. This screen displays the legal cases which have cited the legal case currently being reviewed (i.e., *Pleasant v. Celli* in the example). In this screen shot, the fourth radio button 68 is selected. Thus, the control portion may also have a button 70 which permits the user of the system to limit the types of citations displayed, as described below with reference to Figure 2D. The display portion 60
20 may also display a quotation mark symbol 72 and a depth of treatment symbol 74, which are associated with the citations for the legal case, etc. which cite to the legal case of interest. The quotation symbol 72 indicates that the cited legal case directly quotes from the case of interest (i.e., in the example, *Lubner v. City of Los Angeles* contains a quotation from

Pleasant v. Celli). A method for identifying quotations and verifying the source of the quotations in accordance with the invention will be described below. The depth of treatment symbol 74, which may be, for example, one or more stars, where the number of stars indicate the degree to which the legal case's written opinion is treated, e.g., the amount of text in the cited case opinion which is devoted to the case of interest. The details of the depth of treatment assignment process will be described below in more detail. Now, a screen which permits a user to the limit the citations displayed in the display portion will be described with reference to Figure 2D.

Figure 2D is an example of a screen shot showing the computer screen 50 with the control portion 58 and the display portion 60. In this screen shot, it is assumed the user of the system has selected the limit citation button 70 shown in Figure 2C. As shown, the user of the system may restrict the citations displayed based on headnotes or topics and the system will evaluate all of the citations against the selected headnotes or topics so that only the legal cases containing the selected headnotes or topics are displayed in the screen shot shown in Figure 2C. A headnote may be a few sentences/paragraph which are located at the beginning of a legal case and indicate a summary of the law of a particular portion of the legal case. The user interface of the system permits a researcher to quickly and efficiently perform verification and collocation functions on a legal case. The details of the system for generating information about the legal case and providing the verification and collocation functions in accordance with the invention will now be described.

Figure 3 is a diagram illustrating a method 100 in accordance with the invention, which may be implemented on the computer system of Figure 1, for processing a legal case to generate information about the legal case which may be used for verification and collocation functions. As an aid in understanding the processes, the movement of a single legal case will be described. It should be understood, however, that a plurality of legal cases may be processed at the same time since each legal case may be at a different point in the process. An electronic version of the text of a legal case 102, referred to herein as “WL LOAD”, is fed into a citation identification process 104 (ACITE) that identifies candidate citations to other legal cases and other legal material within the text of the legal case, and marks up the text, i.e., adds a characteristic mark-up symbol to the text, so that the citations may be easily identified at a later time. An example of a mark up symbol may be that the symbol combination “%v” placed at the beginning and at the end of the citation. This identifies the citation for later processing.

Briefly, the citation identification process identifies candidate citations by identifying certain patterns of text in the legal document and compares these patterns to a predetermined set of reference patterns. In particular, digits may be first identified in the text. Next, the text is scanned for abbreviations proximate to the digits which correspond to known reporter abbreviations, such as “Cal.” or “P.”. Once a piece of text having the particular formatting and punctuation of a candidate citation is identified, a case control database 124 is queried to determine if the identified candidate citation corresponds to a valid citation in the case control database. If the identified candidate citation matches a citation in the case control database, a second processing pass is performed. If no match is

located, the identified candidate citation may be flagged for later manual review. As described above, each citation has a predetermined format. The format may be <case name>, <volume number> <abbreviation of reporter name> <series number (if more than one)> <page number in volume>. For example, in “18 Cal. App. 4th 841”. “Cal. App. 4th” refers to the “California Appellate” reporter, 4th series; “18” refers to volume 18; and “841” refers to page 841, the page of volume 18 of Cal. App. 4th where the case decision begins.

As example of a citation to a legal case is *Pleasant v. Celli*, 18 Cal. App. 4th 841, 22 Cal. Rptr 2d 663 (1993) in which the first name portion, i.e., *Pleasant v. Celli*, identifies the parties of the legal case; the second reporter portions, i.e., 18 Cal App. 4th 841 and 22 Cal. Rptr 2d 663, identify the reporters which themselves have a particular characteristic format as described above.

Once text corresponding to a reporter name is located, the text adjacent the reporter name is analyzed to identify the volume, series and page number of the citation as well as the year of the published opinion. Once this information is found, the candidate citation is identified and marked up, as described above, to identify it as a citation. The citation identification process may use a two pass process in which first, full format citations, such as *Pleasant v. Celli*, 18 Cal. App. 4th 841, 22 Cal. Rptr 2d 663 (1993), are identified, matched to the case control database, and placed within a table. In a second pass through the legal case, short form citations, such as *Pleasant*, may be identified based on the text of the full citations that are contained in the table. It should be noted that these short form citations

cannot be identified automatically without first identifying each full citation. For doubtful short form citations which don't match the table, a tentative identification may be made.

The citation identification process 104 in Figure 3 outputs a file 106 containing the text of the legal case with any citations marked up. The file 106 may then be fed into a quote identification process 108 (IQUOTE) in which the text of the legal case is parsed, quotations in the text of the legal case are identified and marked up, and a possible source of the quotation is also identified. At this point, the marked up quotations have not been verified. They are merely candidate quotations which must be further processed to be verified. The details of the quote identification process will be described below with reference to Figures 4-6. The quote identification process may output a file 110 that contains the text of the legal case in which both the citations and the quotations are marked up. At this point, the text of the legal case with the citations and quotations mark-ups may be stored in a database for later use and may also be fed into several processes. These processes may include a quote verification process 112, a depth treatment process 114, and a negative treatment process 116. As shown, these processes may execute in parallel on the same file since each process generates information about the legal case which is separate and independent from that generated by the other processes. Each of these processes will be described in more detail below with reference to Figure 7, Figure 9, and Figure 8, respectively.

In general, the quote verification process 112 verifies that the candidate quotations identified by the quote identification process 108 are in fact from the source (i.e., the citing

case) by comparing the candidate quotation in the cited case to the quotation in the citing case. The process then generates a data record 118 containing information about the verified quotation. The depth treatment process 114 uses information generated by the system, including the verified quotations to generate depth treatment information, such as the number of occurrences of a citation and the characteristics of the citation based on its position (e.g., whether it is free standing, at the head of a string or in the interior of a string).

The process then generates a data record 120 containing information about the depth of treatment information that is applied to each citation in the case of interest. The negative treatment process 116 generates information about any negative treatment the case of interest has received by any of the citing cases and, in step 122, a database 124 containing information about each legal case being processed is updated manually to reflect the negative treatment. The data records 118, 120 from the quotation verification and depth treatment processes, respectively, may be combined together by a grouper process 126 along with a headnote assignment data record 128 (HNRESULT), as described below, to generate a single data record containing the depth treatment information, the quotation information, and the headnote assignment information, about the legal case being processed. This single data record may then be used to generate the information displayed on a computer screen to the user as shown in Figures 2A- 2D.

The data record 118 containing the information about the verified quotations in the legal case also may be fed into a citation loci identification process 130 which attempts to identify the supporting text surrounding quotations and citations in the legal case to generate a citation loci data record 132. The citation loci data record may then be input into a subject

matter assignment and thresholding process 134 which matches the words and phrases in the quotations to one or more headnotes or topics and then determines, based on a threshold value, which headnotes are selected, as described below with reference to Figure 10. The subject matter assignment and thresholding process 134 outputs the data record 128

5 (HNRESULT) containing the selected subject matter text, such as headnotes, which is fed into the grouper 126, as described above. Thus, the system in accordance with the invention automatically generates information about a legal case and then provides that information, using a graphical user interface, to a person using the system when requested. The user may quickly and efficiently locate various information, such as citation information, depth of
10 treatment information, negative treatment information and subject matter text, such as a headnote, about the legal case from a single source. More details about the system will now be described with reference to Figure 4.

Figure 4 is a diagram illustrating more details of the quote identification process 108,
15 the quote verification process 112, the depth treatment process 114, the negative treatment process 116, the citation loci identification process 130 and the subject matter assignment and thresholding process 134 of Figure 3. As shown, the outputs from each of these processes are fed into a system information database 33, as described above.

20 The quote identification process 108 uses the file containing the text of the legal case with marked up citations to identify and mark-up quotations as described above. The text of the legal case contains unverified quotations while the file 144 containing the verified quotations is stored in the database 33. The output of the quote identification process is a

plurality of data records in which each data record has an identified quotation and a possible source of the quotation. The output of the quote identification process may be combined with the file containing the text of the legal case and the marked up citations to produce a file with marked up citations and quotations 110 which is used as an input to the depth of treatment process 114, the negative treatment process 116, the loci identification process 130 and the subject matter assignment and thresholding process 134. During the quote identification process 108, as described below in more detail with reference to Figures 5 and 6, several processes are performed. First, candidate quotations in the text file are identified by scanning the text to identify symbols, e.g., quotation marks, which indicate the beginning or end of a quotation. Next, the beginning and end of the identified quotations are marked up with a quote identifier symbol, such as "%q". Finally, a possible source of the quotations, such as the legal case or other legal material from which the quotations originate is tentatively identified. The source of the quotation is then verified during the quote verification process 112 as described below. The output of the quote identification process 108 may include a Qdata file 140 which contains information about each quotation that is later verified against the probable source of the quotations and a Qtxt file 142 which contains the actual text of the quotations.

The Qdata and Qtxt files 140, 142 are then fed into the quote verification process 112 which uses an electronic database of legal cases, already available, to find and verify the possible source of each quotation found by the quotation identification process. For each quotation, the possible source of the quotation is retrieved. Next, the quotation identified by the quotation identification process is matched against the text of the possible source to

locate text in the source corresponding to the quotation. This verifies the source as the origin of the quotation. For each quotation with a verified source, a data record 144 containing the verified quotations for a legal case is stored in the database 33. Then, when a legal case containing verified quotations is displayed as a citation to a legal case, the citation will contain a quotation symbol, as described above, indicating that the legal case has a verified quotation. The depth treatment process will now be described.

The depth treatment process 114 may receive the file 110 containing the legal case text with the marked up citations and quotations and, in step 146, the depth treatment process performs several processes in order to determine the significance of the citation based on a set of predetermined criteria that are related in some ways to significance. These criteria may be the number of times that the citation appeared in the legal case, the type of the citation, and the association of a verified quotation with the citation. First, the depth treatment process reads through the file 110 and identifies citations which have been marked up previously by the citation identification process. For each identified citation, the type of the citation is determined to be either an ordinary citation, a middle of a string citation, or the head of a string citation. An ordinary citation is a typical citation which usually appears within a legal case and that does not have other citations adjacent to it. A middle of a string (interior) citation is a citation that appears in the middle or at the end of a string citation in which a series of legal documents are cited together in a sentence or paragraph. An interior citation is usually perceived by users as contributing less to the depth with which the cited case is discussed. The head of string citation is a citation that appears at the beginning of a string citation and is perceived by users as contributing more to the depth since it is

conventional to place the most pertinent citation at the head of a string citation. The depth of treatment process may also identify the page number of the legal case for all available pagination on which the citation appears so that a depth record is written as many times as page breaks occur in the legal case.

5

The information about each citation in a legal case, such as the total number of times that the citation appears in the legal case document, the types of each of these citations, and the page number for each citation occurrence is output in a file 148 which is stored in the database 33. This information, in addition to any verified quotations associated with any of the occurrences of the citation, may be used to generate both the “citations to the case” section described above and the depth of treatment symbols. The technique for generating the depth of treatment symbols will be described in more detail below.

10

15

20

The negative treatment process 116 may include an automatic processing step 150 and a manual verification step 152 which generate a list of the negative history (i.e., other written opinions from other legal case which disagree with or overrule the current legal case) for the legal case. During the automatic processing step 150, the file containing the legal text with the marked up citations and quotations is scanned in order to identify stems of certain words, such as “overrule”, “recede”, “disapprove” or “distinguish”, which may indicate negative treatment. As an illustration, the process to identify the root of the word “overrule” in the text of the legal case is described. When an instance of the root “overrule” is identified, a set of heuristic rules, as described below, are applied to make a determination about whether the sentence containing the identified root is actually an overruling, as

described below with reference to Figure 8. Then, during the manual verification process 152, a human operator of the system verifies the results of the automatic process and the actual verified overrulings are added to the case control database 124. The human operator may also identify other negative history about the legal case which cannot be easily identified automatically, as described below. The negative treatment process aids a human operator in rapidly identifying overrulings. These overrulings are negative history which affect the authority of the reasoning of the legal case.

The loci identification process 130 uses the file containing the legal case text with marked up citations and quotations and a file 144 containing the verified quotations, identifies any marked up citations, and applies a set of heuristic rules, as described below, to identify and select a portion of text from around each citation which may indicate the text supported by the citation. If a citation appears multiple times in a legal case, the surrounding text for each of the occurrences of the citation is combined. In addition, if the quote verification process, as described above, has verified any quotation associated with that citation, the text of that verified quotation is also combined with the other text surrounding the citation. All of the identified text that surrounds each citation may then be used to determine one or more headnotes or subject matter headings which may be applicable to the citation. The subject matter heading classifies the citation based on a predetermined number of subject matter areas, such as Intellectual Property or Patents. A process 154 (Headqf) reads all of the text identified adjacent to a given citation and generates a natural language search query to search an existing database for matches to the identified text, as described below. The natural language query process is generally described in U.S. Patent Nos.

5,265,065 and 5,418,948, which are assigned to the same assignee as the present application and are incorporated herein by reference. The Headqf process 154 generates a file 156 containing the natural language queries. Using the natural language queries, a subject matter assignment process step 158 runs the natural language queries against a headnotes database to identify subject matter headings, such as headnotes, which possibly match the text surrounding the citation. For each matched subject matter heading, the query also generates a belief score value indicating how close the subject matter heading match was to the text. A predetermined number of the most closely relevant subject matter headings and their belief scores are provided to a thresholding process step 160.

The thresholding step uses the subject matter headings identified and performs various calculations which take into account the rank of the subject matter headings, the belief score of the subject matter headings and the number of citations which reference that subject matter heading. After the calculations are performed, a predetermined number of top headnote hits and a flag for each headnote indicating if the headnote passed the thresholding are stored in the database 33 with a link to the citation. These subject matter headings permit citations in the legal case to be classified by and searched for using these subject matter headings, as described above with reference to Figure 2D. Now, the quote identification process will be described in more detail.

Figures 5 and 6 are diagram illustrating more details about the quote identification process 108 in accordance with the invention. The quote identification process 108 may include a lexical scanner process 170, a paragraph buffer 172 and a main loop process 174 to

receive the text of the legal case and automatically generate a file containing each quotation identified and a possible source for each quotation. The lexical scanner 170 splits documents into logical fragments, known as tokens, and these tokens are then used by the main loop process 174 to identify quotations. The tokens which are identified by the lexical scanner may include capitalized words, punctuation marks that might end a sentence, white space such as one or more spaces, case names, footnote references, start of quote markers and end of quote markers. The lexical scanner process used may be based on any of a number of commercially available software applications, such as, for example, an application known as FLEX, available from Sun Microsystems Inc, Mountain View, CA.

10 The lexical scanner accepts grammar specifying patterns and identifies an action when a specific pattern is located. In particular, the lexical scanner, in accordance, with the invention may divide a legal case into the certain types of paragraphs based on a predetermined set of criteria, such as a set of rules: 1) a paragraph which might contain a quotation; 2) paragraphs which are indented block quotations; 3) paragraphs which contain

15 important information about the document, such as the start of the document, the document's serial number or the end of the document; and 4) paragraphs which are of no interest to the quotation identification process, such as headnotes, headings and the like. A variety of different criteria and rules may be used to identify these paragraphs.

20 An example of a set of rules which may be used by the invention will now be described. The set may include a rule that identifies paragraphs which do not contain any quotations and stores them in the paragraph buffers where they are overwritten by the next paragraph, and a rule for paragraphs with possible quotations in which the lexical scanner

returns a tag to the main loop indicating that the paragraph is either a normal text paragraph, an indented block quotation paragraph or that the text of the quotation appears in a footnote.

Once the type of the paragraph is determined, the lexical scanner processes the text within the paragraph in the same manner to identify any tokens in each paragraph.

5

Within each paragraph, the lexical scanner may identify the following tokens: a capitalized word, a non-capitalized word, a numeric character string, an abbreviation, a proper name (i.e., "Mr. Smith"), a case citation, a section reference (i.e., "Section 150"), a case name (i.e., "*Roe v. Wade*"), an embedded reference, any end of the sentence

10 punctuation, any other punctuation characters, a colon, semicolon or comma followed by a space, single or multiple white space characters, a start of a quotation, an end of a quotation, the number of a footnote, open and close parentheses, open and close brackets, open and close curly braces, a mark-up for a citation and a mark-up for an embedded reference. More

15 details about the operation and modification of the FLEX software application is available from the Sun Microsystems Inc. Reference Manual, Programmer's Overview Utilities and Libraries, Chapter 9, pp, 203 - 226, which is incorporated herein by reference.

The paragraph buffers 172 are where the tokens about the paragraph most recently scanned by the lexical scanner are stored before being processed by the main loop 174 and

20 then possibly written out into an output file if a quotation is identified in the paragraph. The main loop 174 may decide what action to take for each token returned by the lexical scanner, manage the paragraph buffers and decide when to discard data for a previous paragraph from the paragraph buffer, link several physical paragraphs together into a virtual paragraph for

quotations which run over several physical paragraphs, determine where the breaks between sentences occur within a paragraph, and decide when to process a virtual paragraph by a set of heuristic rules, as described below.

5 Figure 6 is a flowchart of the quotation identification process 108 in accordance with the invention. In step 180, the legal case text is scanned paragraph by paragraph and for each paragraph, the sentences and tokens in the paragraph are identified. In the step 182, a set of heuristic rules is applied to each token in a paragraph to determine if a quotation had been identified. One of the most important functions of the lexical scanner and the quotation
10 identification process is to identify the beginning and end of a quotation. This is difficult since each writer may use a slightly different format for the beginning and ending of a quotation. Therefore, several rules are needed to identify the beginning and ending of a quotation. An example of a set of heuristic rules that may be applied to accomplish such identification will now be described. These rules may use the lexical scanner to identify a
15 conventional start quotation punctuation symbol, such as “ or ‘, to identify a conventional end of quotation delimiter, such as “ or ‘, or to identify a start/end of quotation symbol in a longer string of characters. For example, a rule may attempt to identify strings in which the conventional end of quotation symbol is embedded within a sentence. For each of these rules, the characters surrounding the token may be checked to ensure that the token is in fact
20 a start of end of the quotation.

Once the rules have been applied to each token in a paragraph, the quotation identification process determines if another paragraph exists in step 184 and loop to step 180

to process a new paragraph. Once all of the paragraphs have been analyzed, in step 186, the process output the data record containing the identified quotations and the possible source of those quotations. Now, the quotation verification process will be described.

5 Figure 7 is a flowchart illustrating the method 112 for verifying a quotation in accordance with the invention. At step 200, the quote verification process reads in the text strings identified as quotations by the quote identification process 108 and identifies separators, when present, from a predetermined set of separators in the text strings. The separators may include ellipses, bracketed expressions, and stop phrases. The stop phrases
10 include a variety of legal phrases and others which do not help identify the source of a quotation, for example, "citation(s) omitted", "sic", "emphasis provided" and the like. When present, the separators are used to parse the text string into segments in which each segment includes the words that occur between a pair of separators. In step 202, the text string is parsed to determine its length since the minimum verifiable quote length may be,
15 for example, six non-stop words, where stop words are non-content bearing words such as articles and prepositions. The text string is also parsed to collapse any words which contain apostrophes or other punctuation marks (e.g., "T]hen"). The parsed quotation text string falls into one of two distinct categories: (1) a text string with a single segment, or (2) a text string with multiple segments. Thus, in step 204, the system determines if the text string has
20 a single segment. If the text string has a single segment, then in step 205, the collection normalized inverse document frequency (IDF) for each term (word) in the single segment of the text string is determined. A document frequency value indicates the frequency of a particular term in a typical document collection, while IDF is equal to the reciprocal of

document frequency (i.e., 1/ doc freq), or in other words, the rarity of a term in a document collection. In a preferred embodiment, the collection normalized inverse document frequency (IDF) may be calculated, if the number of occurrences of a word is greater than zero, as:

$$idf_score = \frac{\log\left(\frac{Collection_Docs}{Doc_Occurrences}\right)}{\log(Collection_Docs)}$$

where Doc_Occurrences is the number of documents in which the given term is present and Collection_Docs is the total number of documents in the collection. The IDF is used for purposes of determining good terms for matching, since a rare word is more likely to be distinct and provide a good indication that the quotation is from the candidate source.

Once the IDF has been calculated for each term, a selected number of the terms (i.e., six) with the highest IDF values below a selected threshold may be ranked by IDF value (step 206) and placed into a “template” (i.e., storage array) (step 207) which indicates the position of each term in the text string. Any terms with an unusually high IDF value (e.g., greater than 0.80) are not used, since such infrequently occurring terms are often misspelled words. If there are several terms with the same IDF value, then the alphanumeric ordering of the terms may be used as a secondary key for ranking the terms for the template. Should there still exist equivalent terms (e.g., terms with the same IDF values and alphanumeric spellings) then the position of the terms in the text string may be used as a third key for ranking the terms in the template. The template may then be compared to the quotation from the candidate source document to determine if an exact match, based on the positions of the high IDF terms, occurs in step 208. If an exact match occurs, then in step 210, the verified

quotation is output and fed into the database as described above. In the event that an exact match does not occur in step 212, a certification match failure message is generated and the quotation is not stored in the database.

5 In step 204, if the text string has multiple segments (i.e., it contains one or more separator terms in the text string, such as “The roof fell in... crashing down”), the process goes to step 214 in which the IDF for each term within each required segment is determine. Then, a selected number of terms (e.g., four) within each segment, with the highest IDF values below the threshold, are ranked by IDF (step 215) and placed into a template (step 10 216) in order to determine the position of the terms in the segment for matching purposes (step 208). For a text string with more than four segments, the first two and last two segments may be used to match against the candidate source document (step 217-218). In this manner, the quotations identified by the automatic quote identification process are automatically verified and any verified quotations are identified by a quotation symbol, as 15 described above. Now, the negative treatment process in accordance with the invention will be described.

Figure 8 is a diagram illustrating a method 220 for determining the negative treatment of a legal case in accordance with the invention. The file 110 containing the text 20 of the legal case with the marked up quotations and citations is input into the automatic negative treatment process 150. The automatic negative treatment process may 1) identify occurrences of the word stem “overrul” in the legal case; 2) determine the proximity of the stem to a citation; and 3) exclude any bad legal cases. Prior to identifying the stem

“overrul”, the case control database 124 may be checked and the automatic processing stopped if any history already exists for the legal case. To identify the occurrences of the stem “overrul”, the text of the legal case is are scanned and the verb tense of any occurrences of the stem is determined. The verb tense of the stem indicates whether the overruling refers to the current case overruling a previous case or some other type of overruling. A set of heuristic rules may look for a particular verb tense and then take an action based on the verb tense.

An example of the set of the rules used will now be described, but the invention is not limited to any particular set of rules. For example, one rule may locate “overrule” or “overrules” in a sentence and then scans backwards for up to four words. If “not” or “never” is located, then the sentence is discarded since it does not refer to an actual overruling. If “we” is found, then the sentence is added to the list of possible overruling which are reviewed by a human being. If none of the phrases is located during the backwards scan, the sentence is also added to the list.

Another rule may locate “overrule” and then scans backwards for up to five words to attempt to locate non-case words which would indicate that something other than the legal case is being overruled so that the sentence is not added to the list. A few examples of these non-case words include “request”, “motion”, “objection”, “claim”, and “verdict”. If the rule locates “point” or “points”, then the sentence may be scanned forward to the end of the sentence and if “case”, “cases” or “supra” is located, then the status of the sentence is unknown and it is passed on to the human reviewer.

Another rule may locate “overrule” and scan backwards or forwards , and reject or accepts possible overrulings based on the other words within close proximity to the word “overrule” since these additional words will provide the context in which the word “overrule” is being used. For example, once “overrule” is located, four words before the word may be scanned and the following actions are taken when the following words are located: 1) if “we” is located, and the word prior to “we” is “that”, the “we” is ignored (discussion about overruling only), but if no word “that” is located, then the sentence is a possible overruling; 2) if the verb is modified by a word that indicates uncertainty, such as “rather”, “might”, etc.... the sentence is rejected since the court may be only indicating it might overrule the case: 3) if any word indicates a discussion of an overruling, then the sentence is rejected; 4) if a word indicates that another person did the overruling, then the sentence is rejected; and 5) if “will” or “should” are located, the process looks back five words for a positive word in order to accept the sentence. There may also be a similar set of rules for the verb “overrules”, the infinitive form of the verb and the passive voice of the verb.

Another set of rules may look for various words which indicate a discussion of whether to overrule, whether a court has the authority to overrule or a past overruling since these sentences are rejected as not containing an actual overruling. Another set of rules may reject sentences which indicate that someone else is doing the overruling (i.e., another court in the past). Still another rules may look for “overruling” and then determine if the

sentence is rejected or accepted based on the sentences surrounding the word, as described above.

There are also other rules which look for particular features of a sentence

5 independent of the verb “overrule”. For example, if the phrase “COURT:” is located at the beginning of a sentence, which indicates a direct quotation from the judge, the sentence may be accepted. If the word “Congress” is located at the beginning of a sentence, which may indicate that a Congressional statute is being overruled or that Congress itself is overruling a case, the sentence may be rejected. If the word “circulated” is found in a sentence near the
10 word “overrule”, the sentence may be accepted to catch unusual language, such as “ because the decision overrules an opinion of this court, it was circulated to all active judges...” which could not be automatically identified in some other manner. Another rule may look for “overrule” within a quoted string and reject the sentence since it is usually an overruling by another court of a case which is being quoted by the current court.

15

In addition to the word stem “overrul”, other synonyms may be searched for and identified. For example, the rules may also detect the word stem “abrogat” for California cases which use the term “abrogated” and the phrase “receded from” for Florida cases since these terms are used to indicate an overruling in each respective state. These verb tense rules
20 may be applied in any order and the invention is not limited to any particular set of rules or any particular order of execution of the rules.

The output of the set of verb tense rules from the automatic negative treatment process is a list of possible overrulings. Then, a proximity rule is applied to each possible overruling to determine if the overruling applies to a particular legal case. For example, the proximity rule may eliminate a possible overruling if the sentence containing the stem does not contain a citation, if the previous or next sentence does not contain a citation or the sentence with the stem "overrul" does not contain a word or phrase used to refer to a case, such as "case", "opinion", "holding", "precedent", their plurals or "progeny" or "v.", "ex rel", "ex parte" or "supra". Any sentences which contains the stem "overrul" and satisfies the proximity rules are added to a suggested list 222 of overruling in the legal case. These suggested overrulings are then reviewed and checked during the manual review process step 152 by a human being. The human being, during the manual review process, also determines the case which is overruled and that data is entered into the case control database 124 which tracks legal cases within the legal cases database.

In accordance with another aspect of this negative treatment process, the automatic process may also identify relationships other than overruled, such as "disting" for "distinguished" or "apposite" in a legal case, by extending the method to the language that characterizes those other relationships. In summary, the negative treatment process aids the human reviewer in determining possible overruling in the legal cases by automatically determining possible locations of overruling so that the amount of text that has to be actually reviewed by the human being is significantly reduced. Thus, the negative treatment process increases the speed with which overruling in a legal case may be identified and added into

the negative history of the legal case. Now, the depth treatment process will be described in more detail.

Figure 9 is a flow chart of the depth treatment process 114 in accordance with the invention in which a depth treatment symbol is assigned to each citation within a legal case so that a person using the system may quickly determine the amount of text devoted to discussing a particular citation. This information may be utilized as one indication of the relevance of the citation since a court will devote more text and discussion to a highly relevant citation.

At step 230, the file with the text of the legal case and the marked up citations and quotations, as described above, is received by the depth treatment process. At step 232, the depth treatment process identifies a citation in the legal case, and then in step 234, the type of citation is determined. Each citation in the legal case may be 1) a citation at the head of a string citation; 2) a citation without other accompanying citations; 3) a citation within the interior of a string citation; or 4) a pro forma history citation (i.e., a citation that, in the context of the document, are cited solely as a ancillary historical references for one of the cases cited in its own right). Each of these types of citations has a different amount of significance. For example, a lone citation or a citation at the head of a string citation tends to be more significant than a citation in the middle of the string.

The depth treatment process next determines if there are any additional citations in the legal case in step 236 and loops back to step 232 to process the next citation in the legal

case. Once all of the citations in the legal case have been identified and sorted into one of the types described above, they are fed into the grouper process 126 as shown in Figure 3.

After the grouper process, in step 238, the depth treatment process determines, for each different citation, the total number of each type of citation in the legal case. For example,

5 for a citation to *Pleasant v. Celli*, there may be a total of five cites in the legal case of which three are at the head of a string citation and two are within the interior of a string citation.

This information about each citation in the legal case and any data about a verified quotation which is associated with a particular citation are used in step 240 to determine the depth

10 symbol which will be assigned to the particular citation. Once the depth symbols have been assigned for each citation, the depth treatment process has been completed.

One example of a technique for assigning a depth symbol to a particular citation will now be described, but the invention is not limited to any particular technique for assigning the depth symbols. In addition, the invention is not limited to any particular type of depth

15 symbol. In this example, a citation in the legal case with one to three occurrences of any type of citation (i.e., the citation standing along, the citation is the head of the string citation or the citation is in the middle of a string citation) in the legal case is assigned two stars (e.g.,

), a citation in the legal case with four to eight occurrences of any type of citation is assigned three stars (e.g., *), and a citation with nine or more occurrences of any type of

20 citation is assigned four stars (e.g., ****). To further refine these assignments, a citation with three occurrences of any type of citation and a verified quotation associated with the citation is assigned three stars(e.g., ***) while if a citation has only internal string citation types, one star is deducted from that citation. Thus, the depth symbol for a particular

citation in the legal case is automatically assigned by the system in accordance with the invention. The depth symbols help a user of the system more quickly determine which citations are probably more relevant. Now, the subject matter text assignment process in accordance with the invention will be described.

5

Figure 10 is a flowchart illustrating a method 250 in accordance with the invention for assigning a piece of text from the cited case to the citation in the legal case. In the example described below, the text of a headnote in the cited case is assigned to the citation, but the text assignment process in accordance with the invention may be utilized with a plurality of different pieces of text in the cited cases. In step 252, a citation locus (i.e., a region of text likely to correspond to the text supported by the citation) for each citation is assigned according to the a set of rules which are now described.

10

To identify the citation locus, several text-parsing rules may be used, some of which are stronger than others, but which collectively would be highly likely to identify the text. To allow for varying effectiveness of the different rules, the extracted text may be divided into three groups, “high”, “medium” and “low”, according to the likelihood that the extracted text was part of the correct citation locus. These rules may include:

15

Category	Text
High	<ol style="list-style-type: none"> 1. any non-citation material contained in the sentence that includes the base citation. 2. all of the first non-citation sentence preceding the base citation within the same paragraph. 3. if there is no type 2 sentence, then all of the first non-citation sentence following the citation within the same paragraph. 4. all text that can be identified as a quotation from the cited case.
Medium	<ol style="list-style-type: none"> 5. all sentences occurring between the citation next preceding (but not contiguous to) the base citation and a type 2 sentence. 6. all sentences occurring between the citation next following (but not contiguous to) the base citation and a type 3 sentence. 7. if there is not type 2 or type 3 sentences, and the paragraph containing the base citation ends with a colon or comma, the whole of the paragraph next following the paragraph containing the base citation 8. if there is not type 2 or type 3 sentences and no type 6 paragraph, the whole of the paragraph next preceding the paragraph containing the base citation 9. if any of the text areas identified by any rule includes a headnote reference marker, include the headnote and its keyline
Low	10. the whole of the paragraph containing the base citation
General	11. if the citation occurs in a footnote, it is treated as if it occurred at both the footnote location and the location of the footnote marker
Not Posted	12. citations occurring only in the subsequent history of another citation

Then, in step 254, the terms in the citation loci are weighted according to the rule that was used to identify them, with a high, medium or low matching corresponding to

- 5 weights of 2.0, 1.0 and 0.5, respectively. Different types of documents, such as legal cases or law review articles, may require a different set of rules to determine the weights. Once the pieces of text have been identified and assigned a belief value, in step 256, the identified pieces of text are matched against pieces of text which may be within the cited document. In one example, the pieces of text within the cited document may be headnotes, but the
- 10 invention is not limited to any particular type of text which the identified pieces of text are matched against. The matching may be done using natural language query as described in

previously referenced U.S. Patent Nos. 5,265,065 and 5,418,948 which are owned by the assignee of this application and are incorporated herein by reference. The results of the search is a list of possible pieces of text from the cited case, such as a headnote, which may be assigned to the citation in the legal case and a belief score for each possible piece of text.

5

Next, in step 258, the one or more pieces of text that are going to be assigned to the citation are selected through a thresholding process. The thresholding process ranks the pieces of text for each citation based on the belief score. The piece of text may be posted to the database whenever the following quantity equals or exceeds 0.5:

10

$$\frac{1}{1 + e^{-\left(\beta_0 + \beta_1 \sqrt{\text{belief} - 0.4} + \beta_2 \ln(\text{freq}) + \beta_3 \ln(\text{belief} - \ln 2 + 0.0001)\right)}}$$

The beta values for this equation are as follows:

Document	Rank	β_0	β_1	β_2	β_3
non-ALR	1	4.0451	3.1975	0.8477	0.9033
non-ALR	2	0.5573	9.0220	1.0348	0.6743
non-ALR	3	-2.0421	11.2619	0.8949	0.2954
ALR	1	-1.4256	50.4929	0.3488	0.0000
ALR	2	-2.8199	65.6148	0.6207	0.0000
ALR	3	-2.3701	40.8479	1.2445	0.0000
ALR	4	-3.3474	60.8075	1.7349	0.0000
ALR	5	-3.0805	55.6003	1.3188	0.0000

where the columns marked "ALR" contain variables for ALR articles, as described above, which have a higher belief score than the non-ALR documents. The columns labeled "non-ALR" contain variables for non-ALR documents.

5 In the equation, Freq is the total citation frequency for the citation pair, and lag2 is the belief score of the second following candidate when the candidates are sorted by belief score in descending order (or 0.4 if there is no such candidate). Once the thresholding has been completed and the one or more pieces of text has been assigned to each citation in the legal case, one or more pieces of text are stored in the database in step 260 as described
10 above so that it may be retrieved for a user when requested.

In summary, the subject matter assignment process automatically generates one or more pieces of text for a citation in the legal case based on pieces of text in the cited case, such as a headnote. The process first automatically identifies supporting text in the legal
15 case and assigns a belief value to the supporting text, matches all of the piece of text against pieces of text in the cited cases, and then automatically assigns a piece of text, such as a headnote, from the cited case to the particular citation. These subject matter assignments permit a citation to the legal case to be sorted or selected by the subject matters which helps during the collocation process.

20

Thus, the machine implemented system in accordance with the invention automatically processes a document, such as a legal case, and generates information about the document which may provide the user of the system with useful information about the

contents of the document. In a conventional system, on the other hand, most of this information about the document would be generated by a human being reading the document and making notes about the document which is a slow, expensive, error-prone process. For a legal case, the system may automatically generate information about the negative history of the legal case, about the depth treatment of a citation by the legal case, about the quotation in the legal case which are verified as originating from a particular source, and about one or more headnote which are assigned to a particular citation in the legal case. Thus, the operator of the system may rapidly generate this information about the legal case and a user of the system may quickly locate this information since it is all readily accessible from a graphical user interface.

While the foregoing has been with reference to a particular embodiment of the invention, it will be appreciated by those skilled in the art that changes in this embodiment may be made without departing from the principles and spirit of the invention, the scope of which is defined by the appended claims.